Current Techniques in Health Information Indexing on the WWW

Richard J. Appleyard, PhD

Biomedical Information Communication Center,

Oregon Health Sciences University, Portland, OR, USA

The two main problems facing health information retrieval (IR) on the World-Wide Web (WWW) are (A) the lack of inherent indexing coupled with unparalleled growth, (B) the lack of quality filters coupled with unrestricted ability to publish information. Since there are currently no indexing features built-in to the WWW protocol, most of the efforts to improve health IR have focused on indexing solutions. The significance of context is well-known in traditional health IR, but this problem is even more pronounced on the WWW. This poster seeks to examine the challenges to health information indexing and propose a solution to improving efficiency with contextual indexing.

Two main types of indexing, manual (human) and automated (robot), have been used to organize WWW resources. Manually-maintained hierarchical catalogs, such as the well known Yahoo site [1], were the first approaches used. However, these general catalogs have disadvantages in health IR because they use arbitrary indexing and they are not comprehensive enough. Attempts to maintain comprehensive WWW indexes have used automated software programs called robots (e.g. WebCrawler, Lycos) and powerful computers. However, these free-text indexes are becoming increasingly inefficient when searching for health information due to a "dilution effect". This effect is due to the presence of a large excess of non-medical sites which compounds the problems of word indexing[2], such as polysemy and context.

Medicine-specific manually-maintained catalogs, such as the WWW Virtual Library's Medicine Index [3] and Medical Matrix [4], have provided browsable listings by provider and subject area respectively. More recently, CliniWeb [5] and the Diseases & Disorders Index at the MIC-KIBIC [6] have chosen to catalog clinically-significant sites using the controlled medical vocabulary (Medical Subject Heading, MeSH) developed by the National Library of Medicine. CliniWeb has used this indexing to provide a keyword search capability.

Little headway has been made into the other problems facing health IR on the WWW. In traditional medical information sources, a certain level of quality has been expected and maintained; the medical literature uses the peer-review process and commercial vendors have competitive and legal quality incentives. However, on the WWW there are no such controls since anyone can freely publish anything. Therefore users are concerned about the validity of the information they find. One solution is the reviewing of sites by a third party. Medical Matrix uses a panel to review submissions for quality, and CliniWeb limits indexing to clinically relevant sites. These methods provide somewhat of a quality filter, but fall short of a true peer-review. They also have the potential of excluding other types of useful health information.

Another solution is for the user to be able to critically appraise the source for themselves. Unfortunately, this is not always easy, particularly when a lack of context is provided after traversing a hyperlink to a document buried well within a WWW site. In order to facilitate appraisal, an index of contextual information would be useful so that context could be provided before the hyperlink was traversed.

The context of the WWW document is also important when considering the information needs of the user. Most medical IR research has been performed on bibliographic or medical full-text databases that are used by health care professionals and where the context of the information is clearly defined. This is not the case on the WWW where the medical information available is targeted at all types of user, e.g. practitioner, student or patient, and is created by all types of provider, e.g. hospital, medical school, commercial vendor or individual. Without clearly defined contextual relationships, searches turn up information that may contain the correct search terms, but may be inappropriate to the user.

In order to find out what sorts of contextual labels are typically important when searching the WWW for medical information, protocol analyses were performed on different types of users, i.e. clinicians, students and patients. Users were instructed to "think out aloud" during a WWW search and their actions (screen image and voice) were captured to video tape. Careful analysis of this data allowed the identification of contextual information types used that could be potentially important in WWW health IR. A summary of this information will be presented in this poster. These types are currently being incorporated into a pilot relational database that will store WWW resources focused in a particular medical subject area such as cancer. Searches by different user-types (*vide supra*) using contextual information will be compared with the current free-text word indexing will be used to determine whether this new technique improves search efficiency.

1. Filo, D. and Yang, J., *Yahoo*, <http://www.yahoo.com/>.
2. Hersh, W., *Information Retrieval: A health care prospective.* 1996, NY: Springer-Verlag. 320.
3. Appleyard, R.J., *WWW Virtual Library's Medicine Index*, <http://www.ohsu.edu/cliniweb/wwwvl/>.
4. Malet, G., *Medical Matrix*, <http://www.slackinc.com/matrix/>.
5. Hersh, W., *CliniWeb*, <http://www.ohsu.edu/cliniweb/>.
6. Ahlenius, T., *Diseases & Disorders Index at the Karolinska Institute Medical Information Center (MIC-KIBIC)*, <http://www.mic.ki.se/Diseases/index.html>.

0195—4210/96/$5.00 © 1996 AMIA, Inc.